Samuel Bull

IST 687: Introduction to Data Science

Final Project Report/Response

July 15th, 2020

The data set provided for this project was both extremely interesting and difficult to handle. The processes that I performed throughout my time sitting with this data have helped me to better understand not only the customers that Southeast Airlines has but also the problems and solutions that can come from within the data. At first, the data took a lot to finally format into the what allowed me to come up with my final conclusions and gave me some solutions to what I believe the problems that the airline faces. I also went through and added a few categories that helped me to gain a firmer understanding of ALL the data put in front of me. By creating categories such as ‘Total Money Spent’, by adding up the categories of both ‘Shopping Amount’ and ‘Eating and Drinking’ amount, and ‘Flight Delay in Total’, by adding up the amount of delay in minutes from both departure and arrival of each flight, I was able to examine if the experience of customers both within the terminal and on the flight affected their likelihood to recommend the airline. Speaking of ‘Likelihood to recommend’, I also created logical (which I then converted to numeric) categories for ‘Detractors’ which I utilized later on in my exploration to help work with the airline’s Net Promoter Score as well as work with linear models and Association Rules Mining of the data set. I, also, because I came to the conclusion that it would not affect the data set overall, omitted every row where there was an NA present under any variable.

In terms of basic visualization of the variables that I made numeric within my “airData” data frame, or at least with the ones I saw a benefit in exploring, I created both histograms and boxplots for each category. This, in a way, helped me to come to the conclusion of what an average person (as average as you can get from 5,000 observations) that flies with Southeast is like. In terms of age, both charts show that while people of various ages fly with this airline, the most people who you will find on any given Southeast flight are anywhere between 40 and 45 years old (give or take about 10 years on each side). The people that fly with Southeast are more than likely to fly anywhere between 10-15 times a year, and yet they technically only have an average loyalty of anywhere between -.5 and 0 for Southwest as a company. These people, also on average, are going to spend between $50-$100 while at the airport within their city of origin.

Calling up tables from the non-numeric variables (which I often converted to ordered factors for simplicity sake) also gave insight into who these people are, as well as how they responded to whether or not they would recommend Southeast. In terms of gender, the split was mostly even with slightly more men using the service, with a HEFTY majority of people flying for business reasons as opposed to personal travel or simply using milage points. The price sensitivity of those flying was closer to the lower end of the spectrum as well (1 had the most responses with over 3,300 flyers). Most flights, as would be expected, come from states with major airports, with the California, Florida, Georgia, and Illinois leading the pack with at least more than 300 flights going out within the time period that this data was taken from. The majority of those flying, nearly 4,000 of all total cases actually, flew economy and over 3,300 cases flew on an airline with a status of “Blue”, which makes sense considering the fact that the partner of Southeasts that saw the most action was “CheapSeats Airline Inc.”.

Next, I went ahead and created box plots of what I thought were the most important categorical variables that would help shine some light on why people give their answer to the recommendation question asked. When it came to gender, the outliers for both genders were 2 and 10 but it seemed like more of the men who fly with Southeast were more likely to recommend the airline than that of the women flying (with it looking like the men’s average as 9 while that of the women being closer to a 7). Price sensitivity, as any observer could probably imagine, gave a better outcome when it came to the lower end of the spectrum (the 0s and 1s of the data set) than that of the higher end, with the price sensitivity of 3 having the largest concentrated spread out of all the potential options (where it also had the greatest potential to be a detractor). Those who flew for both business and milage reward-type reasons were likely to recommend the service, while those flying for personal reasons were FAR more likely to be a detractor for the airline. This sentiment translated over into the class boxplot as well, as those flying both economy and economy plus seem to be more likely to be detractors than those who flew business.

From there, I essentially took the same 5 variables used above and created bar charts comparing them to the calculated Net Promoter Score (found using the function nps() in the NPS package I installed). The visualizations that came out of this show essentially the same conclusions I was able to gain from the boxplots I attempted to describe above. Both men and women had positive trends in terms of being a promoter, with men being more likely to recommend Southeast. When it comes to travel type, not only is a person flying for personal reasons far more likely to be a detractor (like was stated previously) but now it shows that, in terms of overall NPS, personal travel actually trends negatively whereas business and milage travel trend positive. Class shows that, while Business is the most likely to be a promoter, economy plus and especially economy flyers have just as much of a chance to be promoters for the airline, and while all airline status’ have the potential to be promoters, those whose airlines were categorized as “Blue” are more likely to trend lower than other status’. Finally, just like in the boxplot above, the price sensitivity has the most probability to be a detractor and actually is the only bar that trends negative in this specific chart.

The generation of a map that captures that average of likelihood to recommend by both origin state and city was the part of this that probably gave me the most trouble. Since there were so many cases within this data set, I had to (and was able to) maneuver my way into finding the average likelihood of both variables I wanted to be able to utilize. However, because the map I utilized only allowed me to map things within the borders of the continental United States, I had to do away with the flights coming from Alaska, Hawaii, and Puerto Rico for the time being. I did this by creating a separate data frame, which I called “airDataMap”, to work with the data that I could plot onto the map. From observing the map I initially created with the average likelihood by state, it seems to me as though the states that were most likely to recommend flying Southeast are closer to the southern mid-west portion of the country while a few exceptions (South Dakota and Maine being the two most obvious ones). With each individual origin city show on the map, it is much harder to tell where exactly the problem with detractors lay, simply because the color scheme that the dots on the map show is hard to distinguish with the naked eye, apart from a few obvious cases. This can immediately be remedied, however, when you separate the obvious promoters and the obvious detractors into their own maps. As was stated before, the most likely origin cities to recommend Southeast are from the southern belt (around the Gulf of Mexico) when you look at the map generated from the separate “promoters” data frame and those that are more likely to be detractors are in Northeastern states and those further out west when looking at the “detractors” map.

Using the “apriori()” function, and messing with both the support and confidence levels that I used in the command (I landed on 0.1 for the support and 0.5 for the confidence), I came up with about 323 rules that could potentially lead to the outcome of a flyer becoming a detractor. After inspecting the ruleset through R, I came to the conclusion that the charts presumptions I made from the charts described above are true. Other than the “Likelihood to recommend” being less than 7, values including the female gender, flying in economy, the airline status being “Blue”, and flying because of personal travel are all factors that could play into someone not recommending the airline.

I also went in and created very basic linear models of what I thought were the 10 most probably variables that would affect a person’s likelihood to recommend Southeast. I was surprised to find that, despite in some cases only having a few categories to base the dependence on (such as when I tested if the gender of a flyer were to affect the likelihood), none of the variables I tested has a high adjusted R-squared value. The highest value I got was when I tested it up against the type of travel people were taking and even then, that translated to only about a 33% dependence. If anything, this shows me that the data collected is far from being extremely black and white in terms of a person’s response. The next closest variable, which made sense to me when I saw the result, was that of an airline’s status (at about 9%), followed by a person’s age (at about 6%) and their gender (at about 1%).

With the support vector machines, I went through and tested a number of variables (mostly the ones that I used to create the tables earlier in the analysis) to try and predict the NPS category (or whether a person would be a promoter, detractor, or passive) and to my surprise, all of the training error and cross validation error values that were returned were much higher than I was expecting them to be (in every single case). I finally settled down on using the “Type.of.Travel” variable to predict the category of each flyer, despite the error margins both being in the 30% range. After running the prediction test on the testing data frame that I created, R returned showing that it predicted 275 detractors correctly and 783 promoters correctly out of the total 1,667 observations in the compare frame that I made in the testing and when calculating the amount of error that occurred because of the test, the result was about the same as when we performed the ksvm() function on the training set.

Overall, the data set that was presented, for lack of a better phrase, was all over the place in terms of what values each observation held. I was able, however, to come up with a number of suggestions from both my observations and the data set in general that I attempt to describe in the presentation that was created alongside this report.